(U) Ask Raul: Numeric Character Reference

FROM: 'Raul', a DNI Analyst
Unknown
Run Date: 11/23/2004

(S) Hey, Raul,
I got this crazy numeric stuff the other day. Here's what it looks like:
&#00068;&#00111;&#00104;&#00033;
Is this some sort of encryption?
Tracy

Dear Tracy,
(U) Believe it or not, that's just text. Some folks refer to it as HTML Escape, or HTML special characters, HTML Unicode or if you want to go by the RFC, Numeric Character Reference (NCR). Here's how it works:

Format:
&# Code ; for decimal (base 10) numbers

&#x Code ; for hexadecimal (base 16) numbers

Code:
Unicode value of the character:

English/Latin letter D = 68 decimal or 44 hex

English/Latin letter a = 97 decimal or 61 hex

Cyrillic letter A = 1040 decimal or 410 hex

Arabic letter Alef = 1575 decimal or 627 hex

Parenthesized Hangul Hieuh = 12813 decimal or 320D hex

Assembled as:

&#68; or &#x44; D

&#1040; or &#x410; Cyrillic A Ð

&#12813; or &#x320D; Parenthesized Hangul Hieuh ã ˆ

It is very simple, is used quite commonly and has been for many years. And it seems to be a "new" thing for most NSA analysts. But it gets better!

(S) Most email tools are normally set by default to send a text and an html version of a mail message. As a result, you can oftentimes take advantage of this little condition to select traffic at the front-end, search for it on the back-end, or recover a perfectly good email from a session which looks lost by simply going after the NCR format.

(S) As you well know, all non-7-bit material going into Lionheart gets indexed as though it were 7-bit. So, a perfectly good foreign text term actually gets indexed as something it is not. If you are lucky, and have a target that uses a character set such as KOI-8, this bit stripping results in what looks like very good translit. If you are unlucky, and your target is using something like CP-1251, then that bit stripping produces a mess. For example, the Russian word for tank, Ñ‚Ð°Ð½Ðº, comes out of this indexing process as ' **TANK** ' for KOI-8 but ' **r`mj** ' for CP-1251.

You're not going to be able to use that CP-1251 term to do anything. However, the same term using NCR would look like this: &#1090;&#1072;&#1085;&#1082; Ñ‚Ð°Ð½Ðº . Aside from being much more unique simply because the pattern contains more characters (bytes), this form is not harmed by the 7-bit indexing procedure.

(S) If you'd like to test this out, go to an Internet terminal or even webworld and type in the string above. If this is the first time you've heard of NCR and you are a linguist, you'll probably be amazed at what you can find on the Internet using NCR strings. There are many things that normally don't show up in your other types of queries.

(S) Making things even better, this NCR form is absolutely beautiful for selection and filtering or searching. What's more, since this form uses nothing but 7-bit characters, it can go through systems, which will destroy 8-bit traffic. This means if you know what you are doing, you can 'recover' a perfect email message from a session, which appears to have been totally lost. Ever seen a piece of email where the entire foreign language portion was nothing but question marks? Chances are that same email had an html double with it containing the NCR. Old Raul used to use this ***magic*** when he was in the Office of Russia to recover "totally destroyed" email. When Raul left, so did the technique. Ð–Ð°Ð»ÑŒ!

(U) So, Tracy, there you go. It's usually the simple little things that bite us in the rear. Oh, if you want a little smile, cut and paste the text below, which includes your sample, into a text editor. Save it as showme.htm and then open it in your web browser.

Enjoy!
Raul

```
<html>
<font size=+4>
&#00068;&#00111;&#00104;&#00033;
</font>
</html>
```